**Using Phylogenetic trees and substitutions to Capture Relationship of** **SARS-Cov-2 Variants and SARS-like Genomes**

Names: Liz Wyman Z1884762

Kleo Bano Z1940978

Chris Troyer Z1945059

Roberto Rivas Z1906735

Emails: Z1884762@students.niu.edu

Z1940978@students.niu.edu

Z1945059@students.niu.edu

Z1906735@students.niu.edu

Main Question: How do variants of SARS-COV-2 (Omicron, Alpha, etc.) and other SARS-like genomes compare to each other?

Background & Significance:

The COVID19 pandemic created a newfound necessity for research in SARS-COV-2 viruses and variants. Since the start of the pandemic more and more variants have been discovered and sequenced. We will use different algorithms to create multiple phylogenetic trees. This will illustrate the different evolutionary relationships between variants and SARS-like genomes. Using multiple algorithms will allow us to create a clear picture of the evolutionary relationship and compare algorithm outputs. Most people do not understand the difference between any of the SARS genomes enough to care about getting an updated shot, so this should help people understand the differences between them to get them thinking about updated covid shots.

Data:

We will use SARS-Cov-2, variants, and SARS-like genomes that can be downloaded from the NCBI website (data files are prepared by Dr. Hou). We will also use the multi-alignment and pairwise alignment data files provided. Each genome has around 30K DNA bases. We will first conduct all vs. all genome comparison and obtain the similarity measurement between every pair of genomes. The similarity is a numeric value between 0 and 1.

The study methods:

A comparison of phylogenetic trees will be used to show differences and similarities between SARS-COV-2, variants, and SARS-like genomes. This will help show why the different genomes and given different names and why SARS-COV-2 variants are variants instead of a different SARS-like genome. We will calculate branch lengths for each algorithm and compare trees. We used distance matrices and BioPython library to get the output on a specific phylogenetic tree. In the case of low substitution rate, we expect the branches to be close together. In the case of high substitution rates, we expect the branches to be far apart from one another. We will first start with the SARS-COV-2 and variants, then move onto SARS-like genomes. Graphs will be constructed and provided as a visual aid to view the data that is found and collected.

Programming language & system: Java, Linux, Python

Library packages: BioJava, Biopython, Matplotlib, ETE

Hours: Liz- 8hours, Kleo- 12hours, Chris-12 hours, Roberto- 8hours

Work environment: GitHub will be used for collaboration on all files and research.

Contribution: Liz – report writing, graph designing, presentation designing 25 %

Kleo - design of the program, implementation/debugging/documentation 25 %

Chris - design of the program, implementation/debugging/documentation 25 %

Roberto - report writing, graph designing, presentation designing 25 %

Initial data (same from proposal):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pw file | Score | Matches | Substitutions | Indels(gaps) |
| sars2.alpha.sing.maf | 2,811,044 | 29,808 | 31 | 10 |
| sars2.beta.sing.maf | 2,842,956 | 29,819 | 26 | 18 |
| sars2.delta.sing.maf | 2,839,244 | 29,761 | 33 | 0 |
| sars2.gamma.sing.maf | 2,836,621 | 29,767 | 35 | 13 |
| sars2.omicron.sing.maf | 2,837,255 | 29,828 | 39 | 36 |
| sars2.sars.sing.maf | 1,589,514 | 23,749 | 5,860 | 369 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Distance matrix based on score** | | | | | | |
|  | **alpha** | **beta** | **delta** | **gamma** | **omicron** | **sars** |
| **alpha** | 0 | 31,912 | 28,200 | 25,577 | 2,611 | 1,221,530 |
| **beta** | 31,912 | 0 | 3,712 | 6,335 | 5,701 | 1,253,442 |
| **delta** | 28,200 | 3,712 | 0 | 2,623 | 1,989 | 1,249,730 |
| **gamma** | 25,577 | 6,335 | 2,623 | 0 | 634 | 1,247,107 |
| **omicron** | 26,211 | 5,701 | 1,989 | 634 | 0 | 1,247,741 |
| **sars** | 1,221,530 | 1,253,442 | 1,249,730 | 1,247,107 | 1,247,741 | 0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Distance matrix based on matches to sars 2** | | | | | | |
|  | **alpha** | **beta** | **delta** | **gamma** | **omicron** | **sars** |
| **alpha** | 0 | 11 | 47 | 41 | 20 | 6,059 |
| **beta** | 11 | 0 | 58 | 52 | 9 | 6,070 |
| **delta** | 47 | 58 | 0 | 6 | 67 | 6,012 |
| **gamma** | 41 | 52 | 6 | 0 | 61 | 6,018 |
| **omicron** | 20 | 9 | 67 | 61 | 0 | 6,079 |
| **sars** | 6,059 | 6,070 | 6,012 | 6,018 | 6,079 | 0 |
| **Distance matrix using substitution rate** | | | | | | |
|  | **alpha** | **beta** | **delta** | **gamma** | **omicron** | **sars** |
| **alpha** | 0 | 1.68 X 10-4 | 6.9 X 10-5 | 1.36 X 10-4 | 2.67 X 10-4 | 1.97 X 10-1 |
| **beta** | 1.68 X 10-4 | 0 | 2.36 X 10-4 | 3.03 X 10-4 | 4.35 X 10-4 | 1.97 X 10-1 |
| **delta** | 6.9 X 10-5 | 2.36 X 10-4 | 0 | 6.7 X 10-5 | 1.98 X 10-4 | 1.97 X 10-1 |
| **gamma** | 1.36 X 10-4 | 3.03 X 10-4 | 6.7 X 10-5 | 0 | 1.31 X 10-4 | 1.97 X 10-1 |
| **omicron** | 2.67 X 10-4 | 4.35 X 10-4 | 1.98 X 10-4 | 1.31 X 10-4 | 0 | 1.97 X 10-1 |
| **sars** | 1.97 X 10-1 | 1.97 X 10-1 | 1.97 X 10-1 | 1.97 X 10-1 | 1.97 X 10-1 | 0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Distance matrix using gap rates** | | | | | | |
|  | **alpha** | **beta** | **delta** | **gamma** | **omicron** | **sars** |
| **alpha** | 0 | 2.68 X 10-4 | 3.35 X 10-4 | 1.01 X 10-4 | 8.69 X 10-4 | 1.20 X 10-2 |
| **beta** | 2.68 X 10-4 | 0 | 6.03 X 10-4 | 1.67 X 10-4 | 6.01 X 10-4 | 1.17 X 10-2 |
| **delta** | 3.35 X 10-4 | 6.03 X 10-4 | 0 | 4.36 X 10-4 | 1.20 X 10-3 | 1.23 X 10-2 |
| **gamma** | 1.01 X 10-4 | 1.67 X 10-4 | 4.36 X 10-4 | 0 | 7.68 X 10-4 | 1.19 X 10-2 |
| **omicron** | 8.69 X 10-4 | 6.01 X 10-4 | 1.20 X 10-3 | 7.68 X 10-4 | 0 | 1.11 X 10-2 |
| **sars** | 1.20 X 10-2 | 1.17 X 10-2 | 1.23 X 10-2 | 1.19 X 10-2 | 1.11 X 10-2 | 0 |

Distance matrix analysis:

There are 4 different distance matrices above that show how closely related SARS-COV-2 variants and sars are to one another. Comparing the data from all different matrices shows the substantial differences between SARS-COV-2 variants and sars. The variants differences to each other vary within a smaller range of numbers, suggesting some variants are more closely related than others. This is useful in determining precautions to use against each variant.

Substitution rate analysis:

Comparing the wildtype SARS-COV-2 genome to SARS-COV-2 variants and SARS-like genome sars, shows how the substitution rate is substantially greater for sars than any of the variants. Also, the substitution rates comparing sars to variants are within a ~0.002 margin to the substitution rate of sars to SARS-COV-2 of 0.194913. This shows how closely related the variants are compared to one another and the wildtype SARS-COV-2 genome, which is why they

are not considered variants of SARS-like genome sars.

Phylogenetic tree analysis:

As you can see in the tree below, the omicron variants have longer branch lengths and are clustered together. This shows how closely related the omicron variants are, while showing how they differ from other SARS-COV-2 variants. The only SARS-COV-2 variant that has a relatively similar branch length is eta. Although, more recent omicron variants have longer branches. A branch indicates evolutionary change from the root sars2 and as the branch length increases it shows the amount of evolutionary change also increases. This also suggets how a covid shot that targeted the alpha variant would not be as effective at targeting omicron variants because of how much they changed. More and more variants keep evolving, so it is important to keep an eye on their level of relation to one another to better understand how to take precautionary measures against them.

A diagram of a family tree

Description automatically generated

SARS-COV-2 variant tree analysis:

A black and white chart

Description automatically generated with medium confidenceLooking at the tree below that removes omicron variants from the equation can give a better look at how the SARS-COV-2 variants relate to one another. Epsilon shows the least amount of evolutionary change and eta shows the greatest when comparing only SARS-COV-2 variants. This is determined by considering the branch length. The longer the branch length the more distant an evolutionary relationship.

UPGMA Trees analysis:

These trees are distance-based trees in which leaves that share the same root have the same distance. The mutation rate is considered a constant throughout the entire tree, which is why the SARS-COV-2 variants branch distances are much more uniform than in the Parsimony branches. UPGMA trees are less reliable than other trees because of the mutation rate considered as constant.A black and white image of a tree

Description automatically generated

A computer screen shot of a computer

Description automatically generatedA black and white image of a diagram

Description automatically generated

A white background with black text

Description automatically generated

Parsimony trees analysis:

These trees are character-based trees that seek the lowest number of mutations possible for the simplest trees. Parsimony trees on their own give an incomplete picture because of reliance on one method of analysis, but when it is similar to other methods it gives a more complete picture of reliance. These parsimony trees look similar to the UPGMA trees because the algorithm for parsimony uses UPGMA as a base to start. Additionally, the parsimony trees used Fitch’s algorthim for simplicity as Sankoff’s algorithm is more complicated and takes additional time to complete.

A white background with black lines

Description automatically generatedA computer screen shot of a computer code

Description automatically generated with medium confidence

A white paper with black text

Description automatically generatedA computer screen shot of a diagram

Description automatically generated

Neighbor-Joining trees analysis:

These trees are distance-based trees that find a pair of leaves close to each other and far from other leaves. This creates an unrooted tree and considers variations on mutation rates, which make it more reliable than trees like UPGMA. As you can see with these trees too the omnicron variants seem to have their own branch and a re more closely related to each other than any other variant.

A circular pattern of black and white text

Description automatically generated

A spiral of black and white text

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

A diagram of a sun

Description automatically generated with medium confidence